Multilingual Non-Native Speech Recognition using Phonetic Confusion-Based Acoustic Model Modification and Graphemic Constraints

Ghazi Bouselmi, Dominique Fohr, Irina Illina, Jean-Paul Haton

Multilingual Non-Native Speech Recognition using Phonetic Confusion-Based Acoustic Model Modification and Graphemic Constraints

G. Bouselmi, D. Fohr, I. Illina, J.-P. Haton

Speech Group, LORIA-CNRS & INRIA, "http://parole.loria.fr/"
BP 239, 54600 Vandoeuvre-ls-Nancy, France
{ bousselm,fohr,illina,jph }@loria.fr

Abstract

In this paper we present an automated approach for non-native speech recognition. We introduce a new phonetic confusion concept that associates sequences of native language (NL) phones to spoken language (SL) phones. Phonetic confusion rules are automatically extracted from a non-native speech database for a given NL and SL using both NL's and SL's ASR systems. These rules are used to modify the acoustic models (HMMs) of SL's ASR by adding acoustic models of NL's phones according to these rules. As pronunciation errors that non-native speakers produce depend on the writing of the words, we have also used graphemic constraints in the phonetic confusion extraction process. In the lexicon, the phones in words' pronunciations are linked to the corresponding graphemes (characters) of the word. In this way, the phonetic confusion is established between couples of (SL phones, graphemes) and sequences of NL phones. We evaluated our approach on French, Italian, Spanish and Greek non-native speech databases. The spoken language is English. The modified ASR system achieved significant improvements ranging from 20.3% to 43.2% (relative) in *sentence error rate* and from 26.6% to 50.0% in *WER*.

Index Terms: non-native speech recognition, pronunciation modelling, graphemic constraints.

1. Introduction

The performance of automatic speech recognition (ASR) systems drastically drops when confronted with non-native speech. Classical ASR systems are trained with native speakers and designed to recognize native speech. The statistical methods they are based upon do not handle pronunciation variants or accents that non-native speakers produce.

Non-native speech enhancement of existing ASR systems aims at making those systems more tolerant to pronunciation variants and accents produced by non-native speakers. Several approaches have been developed in that respect. They differ in the techniques used to extract the knowledge about pronunciation variants and to integrate them into the ASR system. In [3], this knowledge is extracted by human experts with a study of phonological properties of the NL and SL. A set of phone rewriting rules is specified for each spoken/native language pair. These rules are then used to modify the lexicon of the ASR. In [4], authors used a non-native speech database in order to automatically extract a phonetic confusion matrix : the canonical pronunciation (SL phones) and the actual one (SL phones) are aligned for each utterance. The lexicon is then dynamically modified to include all possible pronunciations during the recognition phase. In [5], a confusion matrix is established between the SL's and NL's phones. The SL's ASR system is used to align the canonical pronunciation of each utterance. The NL's ASR system supplies an actual pronunciation with NL's phones for each utterance, using phonetic recognition. The two transcriptions are aligned in order to extract the phonetic confusion. Finally, the Gaussian mixtures of the acoustic models of each NL's phone are merged with those of the SL's phones they were confused with. These new models are then used in the modified ASR system.

As we studied non-native speech, we have spotted two main problems that ASR systems are faced with.

First, we noticed that non-native speakers tend to pronounce phones as they would do in their native language. Phones of the SL are often pronounced as similar phones from the NL. Phones of the SL that do not exist in the NL are an obvious example. For instance, the English phone '[ð]' (present in the word "*the*") is often pronounced as the French phone '[z]' by French speakers. Furthermore, some SL phones may correspond to a sequence of NL phones as for the English phone '[aɪ]' that may be pronounced as the sequence of French phones '[a] [ɪ]'. Thus, we introduced a new approach for phonetic confusion in [1]. This confusion associates sequences of NL's phones to each SL's phone. The SL phone models are modified according to this confusion.

Second, we noticed that the writing of uttered words influences the pronunciations produced by non-native speakers. The pronunciation errors made by non-native speakers are closely related to the writing of words. The same phone is pronounced differently according to the character it is related to in the word. Furthermore, when faced with difficult or unknown pronunciations, non-native speakers utter words in a similar manner to their mother tongue. Let's consider the example of the table 1 where the canonical pronunciation and actual pronunciation made by a French speaker are illustrated for the English words "*approach*" and "*position*". The English phone '[ə]' is pronounced by some French speakers as the French phone '[ɔ]' when it corresponds to the character 'o' and as the French phone '[a]' when it corresponds to the character 'a'. We suppose that taking into account the writing of the words may further enhance the performance of the speech recognition. Thus, we have introduced graphemic constraints in the phonetic confusion in [2]. These graphemic constraints are used to modify the lexicon of the ASR system.

In this paper, an extended evaluation of these two methods and a comparison with MLLR adaptation are presented. The

Table 1: *Phonetic transcription and actual pronunciation of the English words "approach" and "position" by one French speaker.*

Word	*"Approach"*	*"Position"*
Canonical transcription	**[ə]** [p] [r] [əʊ] [tʃ]	[p] **[ə]** [z] [i] [ʃ] **[ə]** [n]
Actual pronunciation	**[a]** [p] [r] [ɔ] [t] [ʃ]	[p] **[ɔ]** [z] [i] [ʃ] **[ɔ]** [n]

database used is composed of English speech uttered by French, Spanish, Greek and Italian speakers.

In the next sections, the phonetic confusion concept and the graphemic constraints are described. Then, several test results are presented. Finally, these results are discussed in a brief conclusion.

2. Brief overview

As described in figures 1 and 2, the SL ASR system, the NL ASR system and a non-native speech database are used to extract the phonetic confusion and modify the target ASR system (SL ASR system). The graphemic constraints may be applied to the ASR system prior to the application of the phonetic confusion. Applying the graphemic constraints to an ASR system consists in linking the phones to the character in the word pronunciations and modifying the lexicon accordingly.

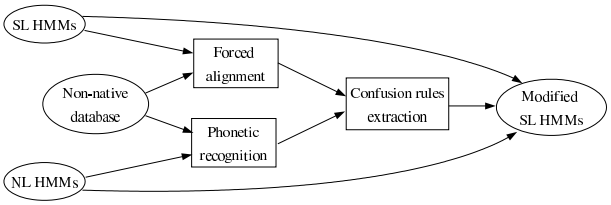

Figure 1: *Extracting and using the phonetic confusion.*

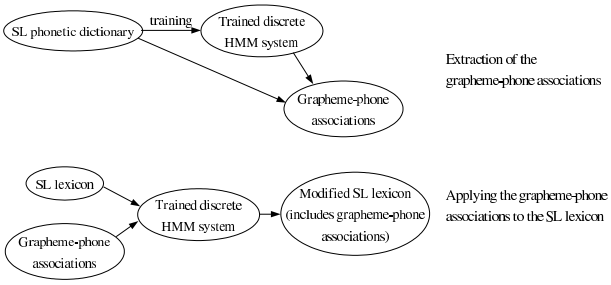

Figure 2: *Extracting and applying the graphemic constraints.*

3. Inter-language phonetic confusion

We will briefly recall our phonetic confusion concept described in [1]. Non-native speakers tend to pronounce phones as in their mother tongue. Besides, in some cases, phones of the SL may not exist in the NL or may correspond to a sequence of NL's phones. Thus, the phonetic confusion we developed involves phones of both the SL and the NL. Phones of the SL are associated with sequences of phones of the NL.

3.1. Extracting the phonetic confusion

As stated above, we use both SL's and NL's ASR systems to extract the phonetic confusion. The SL's ASR system is used to perform a phonetic alignment of the canonical pronunciation for each utterance of the non-native database. The NL's ASR system supplies a phonetic transcription in terms of NL's phones for these utterances (by a phonetic recognition). By aligning those two transcriptions for each utterance, we extract associations between SL's phones and sequences of NL's phones. A SL phone $[K]$ is associated with the NL phones sequence $(M_i)_{i\in I}$ if all phones M_i have at least half of their time interval included in $[K]$'s one.

The next step is to extract the phonetic confusion rules from these associations. The maximum likelihood of the probability of each phone association is computed ($P(K ==> (M_i)_{i\in I})$) for each phone $[K]$. Only the most probable associations are retained to make up the confusion rules set.

Here is an example of the rules, extracted by our system, for the English phone '[aɪ]'. NL is Italian :
[aɪ]==>[a] [ɪ] $P([aɪ]==>[a]\ [e]) = 0.4$
[aɪ]==>[a] [ɪ] $P([aɪ]==>[a]\ [ɪ]) = 0.6$

3.2. Using the phonetic confusion

The acoustic models of the SL's ASR system are modified using the phonetic confusions extracted in the previous step. For each SL phone $[K]$, HMMs of the sequences of NL phones that were confused with $[K]$ are added as alternative paths to the HMM of $[K]$. Assuming the rules sketched in section 3.1, the figure 3 illustrates the construction of the modified HMM for the English phone [aɪ]. In the figure 3, β is a weight between the NL phones and SL phones.

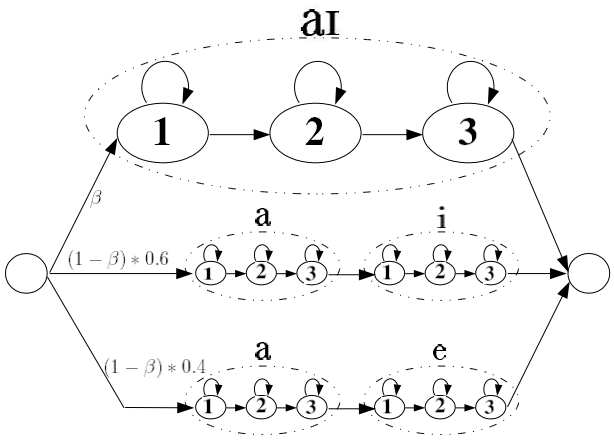

Figure 3: *Modified HMM model structure for English phone [aɪ].*

This way, the excessive computational overload that results in modifying the lexicon is avoided : as stated in [4], adding all the possible pronunciations to the lexicon leads to an excessive growth of the lexicon. Furthermore, the coherence of the acoustic models is preserved as opposed to GMM merging in [5].

4. Graphemic constraints

We assume that taking into account the writing of the words in the phonetic confusion extraction process may enhance the performance of the modified ASR system. In this step, we automatically associate the phones to the characters they are related to in the pronunciation of the words of the lexicon. Graphemic constraints have already been used in non-native ASR system enhancement. Nevertheless, the existing approaches do not use an automated process to perform the grapheme-phone alignment (as in [3]).

4.1. Automatic grapheme-phone alignment

Given the writing of a SL word and its pronunciation, the goal is to associate each phone of SL to the graphemes of SL (characters) they are related to. The task of grapheme-phone alignment differs from grapheme to phone translation. Rather, the knowledge that we seek is the link between graphemes and phones in each word pronunciation.

4.1.1. Extracting the graphemic constraints

The grapheme-phone alignment is automatically extracted from a phonetic dictionary. The phonetic dictionary is used to train discrete HMMs. In this system, graphemes represent the discrete observations, phones represent the HMM models.

The initial discrete HMM models have a uniform emission probability among all discrete symbols (one symbol for each grapheme). The system is then trained on the phonetic dictionary in order to learn the grapheme-phone associations. The next step consists in extracting the explicit grapheme-phone associations. The trained discrete HMM system is used to perform a forced alignment on the training dictionary. For each word of this dictionary, the phones (representing the discrete HMMs) are associated with the character(s) (representing the observations) according to the result of the alignment. For each phone, only the most often encountered grapheme-phone associations are retained. An association a_K for a phone $[K]$ is kept only if satisfies the equation (1) :

$$N(a_K) \geq \gamma \, \Sigma_{a'_K \in A_K} N(a'_K) \quad (1)$$

where A_K is the set of grapheme-phone associations for phone $[K]$, $N(a_K)$ is the count of appearance of the association a_K, and γ is a factor.

4.1.2. Applying the graphemic constraints to the ASR system

We propose a straight forward approach to integrate the graphemic constraints in the target ASR system. We modify the lexicon by replacing each phone by the couple of (phone, grapheme) related in the pronunciation of each word. Word pronunciations are no longer a sequence of phones. Pronunciations will consist in sequence of couples of (phone, grapheme). Here is an example for the English word "*speech*":

phonetic transcription	[s] [p] [i:] [tʃ]
grapheme-phone association	([s], S) ([p], P) ([i:], EE) ([tʃ], CH)

To achieve this modification, the trained discrete HMM system is used. A forced alignment is performed on the dictionary of the target ASR system using the discrete HMM system. We obtain the grapheme-phone associations for each phone in the pronunciation of the words (in the target dictionary). The pronunciation of each word in the target lexicon is modified according to these associations. Only the associations that appear in the set extracted from the training dictionary are retained (see previous section). If an association is not retained for a phone $[K]$ in a word W, the phone $[K]$ remains without graphemic constraint in the pronunciation of W.

The last modification consists in adding HMM models for the newly introduced phones (in the target ASR system). For each added phone $[K]$ with a graphemic constraint X, a new HMM model ($[K]$, X) is added to the system. The model for the phone ($[K]$, X) is a copy of the model for the phone $[K]$, since, it is the same phone.

4.2. Alignment issues

Using a discrete HMMs system has raised a problem in the grapheme-phone alignment. For example, the grapheme-phone alignment for the English word "*used*" requires some phones to share the same grapheme. This word is pronounced [j] [u:] [z] [d]. The straight forward application of the grapheme-phone method above will lead to the following wrong result: ([j], U), ([u:], S), ([z], E) and ([d], D). We have chosen to duplicate the observations that the discrete HMM system processes. For example, for the word "*used*", the discrete system will process the sequence (U, U, S, S, E, E, D, D) rather than (U, S, E, D). We introduce this data duplication in order to get the following alignment for the word "*used*": ([j], U), ([u:], U), ([z], SS), ([d], EEDD). A post-processing will lead to the correct alignment: ([j], U), ([u:], U), ([z], S), ([d], ED).

5. Experiments

The work presented in this paper has been done in the framework of the European project *HIWIRE* which aims at enhancing ASR in mobile and noisy environments. The *HIWIRE* project deals with the development of an automatic system for the control of aircrafts by pilots via voice.

5.1. Experimental conditions

We used a non-native speech database consisting of 31 French speakers, 20 Italian speakers, 20 Greek speakers and 10 Spanish speakers. Each on of these speakers utters 100 English sentences (a random list), read speech and noise-free recording. The sentenses are composed of an average of 3-4 words. The acoustic parameters are 13 MFCCs with their first and second time derivatives. The 46 English monophone models have been trained on the *TIMIT* database. The French, Italian, Greek and Spanish monophone models have been trained respectively on French, Italian, Greek and Spanish native speech databases. The HMM models have 128 Gaussian mixtures per state and diagonal covariance matrices. We used the toolkit HTK in order to train the models, the decoder is a time-synchronous viterbi decoder. The vocabulary is composed of 134 words. The grammar is a command language (*strict grammar*) and a "word-loop grammar". The development set consists in the 50 first sentenses from all speakers of same native language. A global phonetic (speaker independent) confusion is extracted using the development set for each native language

Table 2: *Test results on the French, Italian, Spanish and Greek databases (in %).*

	French		*Italian*		*Spanish*		*Greek*		*Average*	
System	*WER*	*SER*	*WER*	*SER*	*WER*	*SER*	*WER*	*SER*	*WER*	*SER*
strict grammar:										
- baseline	6.0	12.8	10.5	19.6	7.0	14.9	5.8	13.2	7.3	15.1
- "phonetic confusion"	**4.4**	**10.2**	**6.9**	**14.1**	**5.1**	**11.8**	**2.9**	**7.5**	**4.8**	**10.9**
- "phonetic confusion" + graphemic constraints	4.9	11.3	8.2	15.9	6.2	13.6	6.0	15.1	6.3	14.0
- baseline + MLLR	4.3	8.9	7.3	13.6	5.1	11.1	3.6	9.4	5.1	10.8
- "phonetic confusion" + MLLR	**3.1**	**7.2**	**4.9**	**11.5**	**3.4**	**8.0**	**2.3**	**6.5**	**3.4**	**8.3**
- "phonetic confusion" + graphemic const. + MLLR	3.7	8.5	6.5	14.1	4.8	9.8	4.8	12.7	5.0	11.3
word-loop grammar:										
- baseline	37.7	47.9	45.5	52.0	39.9	53.5	36.7	40.0	40.0	50.7
- "phonetic confusion"	27.3	42.1	31.3	46.2	**29.5**	**44.5**	**20.3**	**35.1**	**27.1**	**42.0**
- "phonetic confusion" + graphemic constraints	**26.2**	**41.9**	**30.5**	**45.5**	31.3	46.5	24.3	43.0	28.1	44.2
- baseline + MLLR	28.4	39.4	34.9	46.5	32.3	48.3	28.5	41.0	32.2	42.7
- "phonetic confusion" + MLLR	**23.0**	**36.6**	**25.2**	**40.6**	**24.7**	40.1	**18.1**	**31.3**	**22.8**	**37.2**
- "phonetic confusion" + graphemic const. + MLLR	**23.0**	**36.6**	25.6	41.2	25.9	**39.6**	20.8	37.5	24.1	39.0

group. The speech recognition tests were performed on the 50 last sentences of each speaker (test set).

5.2. Results

We tested the baseline system (English ASR without modifications), the phonetic confusion, the phonetic confusion along with the graphemic constraints and MLLR speaker adaptation. We carried out separate tests on the French, Italian, Greek and Spanish. Phonetic confusion rules have been extracted for each native language using the proper acoustic models and the development set. We have limited the phonetic confusion to only 2 confused phone sequences in all the tests.

The table 2 summarizes the results of the different tests. In comparison to the baseline system, the phonetic confusion approach achieved significant improvements varying between 20.3% and 43.2% (relative) in *sentence error rate* (SER) and between 26.6% and 50.0% (relative) in *word error rate* (WER). Using the word-loop grammar, these improvement range from 11.2% to 29.1% (relative) in SER and from 21.6% to 45.0% (relative) in WER. As shown in table 2, the use of phonetic confusion outperforms the MLLR speaker adaptation when using a strict grammar. Using the strict grammar, the use of the graphemic constraints did not lead to an improvement as compared to the phonetic confusion alone. Nonetheless, the graphemic constraints with the phonetic confusion allowed slight improvements over the confusion alone when using the word-loop grammar (except for the Greek database). We think that the grammar used in our application, a strict command language grammar, makes further improvements difficult to achieve, especially when using the graphemic constraints. Besides, the small size of our databases prevents the extraction of reliable phonetic confusion rules with graphemic constraints.

6. Conclusion

In this paper we presented extended evaluation for several native languages of our approach for non-native speech recognition. This approach is based on a new phonetic confusion concept and the graphemic constraints. The experiments are carried on database of English speech uttered by French, Spanish, Greek and Italian speakers. The use of phonetic confusion lead to significant improvements in recognition rates for all four languages compared to the MLLR adapted system. On the other hand, the use of graphemic constraints gives a slight improvement while using a word loop grammar. The use of the MLLR speaker adaptation after phonetic confusion-based acoustic model modification allowed further improvements.

7. Acknowledgments

This work was partially funded by the European project *HIWIRE* (*Human Input that Works In Real Environments*), contract number 507943, *sixth framework program, information society technologies*.

8. References

[1] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton, "Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration". In Proc. Eurospeech/Interspeech, Lisboa, September 2005.

[2] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton, "Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration and Graphemic Constraints". In Proc. ICASSP, Toulouse, France, May 2006.

[3] Stefan Schaden, "Generating Non-Native Pronunciation Lexicons by Phonological Rule". In Proc. ICSLP 2004.

[4] K. Livescu and J. Glass, "Lexical Modeling of Non-Native Speech for Automatic Speech Recognition", In Proc. ICASSP, 2000.

[5] J. Morgan, "Making a Speech Recognizer Tolerate Non-Native Speech Through Gaussian Mixture Merging". In Proc. InSTIL/ICALL 2004.